SP²RINT: Spatially-Decoupled Physics-Constrained ProgRessive INverse OpTimization for Diffractive Optical Neural Network Training

Pingchuan Ma
pingchua@asu.edu
Arizona State University
Tempe, Arizona, USA

Ziang Yin
ziangyin@asu.edu
Arizona State University
Tempe, Arizona, USA

Qi Jing
qjing1@asu.edu
Arizona State University
Tempe, Arizona, USA

Zhengqi Gao
zhengqi@mit.edu
Massachusetts Institute of Technology
Boston, Massachusetts, USA

Nicholas Gangi
gangin2@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York, USA

Boyang Zhang
bzhang523@wisc.edu
University of Wisconsin at Madison
Madison, Wisconsin, USA

Tsung-Wei Huang
tsung-wei.huang@wisc.edu
University of Wisconsin at Madison
Madison, Wisconsin, USA

Rena Huang
huangz3@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York, USA

Duane S. Boning
boning@mtl.mit.edu
Massachusetts Institute of Technology
Boston, Massachusetts, USA

Yu Yao
yuyao@asu.edu
Arizona State University
Tempe, Arizona, USA

Jiaqi Gu
jiaqigu@asu.edu
Arizona State University
Tempe, Arizona, USA

Abstract

Diffractive Optical Neural Networks (DONNs) harness the physics of light propagation to perform speed-of-light information processing. Training DONN systems to determine the metasurface structures remains a challenging problem. Heuristic methods are fast but oversimplify metasurfaces as element-wise phase masks or convolution units, often resulting in physically unrealizable designs. Simulation-in-the-loop training methods directly optimize metasurfaces using adjoint methods during training, but are computationally prohibitive. To address the DONN training in a physically feasible and scalable manner, we propose a spatially decoupled, progressive training scheme, SP²RINT. For the first time, we formulate DONN training as a partial differential equation (PDE)-constrained learning problem, where metasurface responses are relaxed into trainable, banded transfer matrices. We then progressively enforce physical constraints through alternating transfer matrix training and inverse design. It eliminates the need for costly PDE solving per step while ensuring physical realizability. To further alleviate the runtime bottleneck, we partition the metasurface into independently solvable patches and optimize the transfer matrix of each sub-region in parallel, followed by system calibration to restore global consistency. Across a range of DONN training tasks, SP²RINT achieves digital-comparable accuracy while being 1825× faster than simulation-in-the-loop approaches. Grounded in a physics-inspired optimization, SP²RINT bridges the gap between abstract DONN models and physical metasurface designs and paves the way for scalable training of DONNs with guaranteed physical feasibility and high accuracy. We open-source our code at SP²RINT.

1 Introduction

Diffractive optical neural networks (DONNs) have emerged as a promising technology, leveraging the inherent parallelism for analog processing at light-speed in diverse tasks such as computer vision, AI inference, scientific computing, sensing, and imaging [3, 5, 12–16, 18, 22, 24–26, 35, 41]. One emerging technology to implement compact, flat optical field manipulation in DONNs is metasurfaces. Metasurfaces are engineered planar optical elements, composed of arrays of subwavelength-scale structures known as meta-atoms.

However, training DONNs and designing physical metasurfaces remain notable challenges, rooted in the complexities of physical constraints. Traditional DONN design methods fall into two categories: *heuristic* approximation methods and *simulation-in-the-loop* training [11]. Heuristic methods simplify the metasurface as a set of element-wise phase masks or convolution kernels, often bypassing rigorous simulations. These methods translate the desired phase profile into a metasurface layout using a pre-simulated look-up table (LUT) that maps each target phase to a corresponding meta-atom design. Though efficient, this approach often neglects important inter-element interactions and relies heavily on the local periodic approximation (LPA) [35]. The LPA assumes a smooth near-field response, a condition that might hold in some lens designs but is frequently violated in DONNs where strong spatial variations are common. As a result, heuristic methods often produce target phase profiles or transfer matrices that are not physically realizable. When these idealized designs are projected onto feasible metasurface structures, the mismatch leads to significant deviations in optical behavior and degraded DONN performance. Efforts to mitigate this issue, e.g., by smooth phase regularization [14, 31, 41] to ease hardware design, have generally been inadequate, as they significantly limit DONN expressivity and do not ensure true physical feasibility, often leading to impractical designs.

In contrast, simulation-in-the-loop training[11] embeds metasurface optimization directly in the training loop by leveraging adjoint methods. This approach enforces physical feasibility at every training iteration. However, the demanding requirement of solving forward/adjoint Maxwell's equations for each metasurface at each iteration makes it prohibitively expensive and fundamentally unscalable.

To address these bottlenecks in DONN training, we propose SP²RINT, a physics-constrained optimization scheme that progressively enforces Maxwell PDE constraints during training. Unlike prior heuristic or simulation-in-the-loop approaches, SP²RINT alternates between

relaxed DONN training and inverse metasurface projection, ensuring that learned responses remain physically realizable while maintaining efficient design space exploration.

To further reduce the significant computational cost, we develop a patch-wise simulation strategy. This strategy effectively exploits the locality of meta-atom interactions and inherent spatial frequency limits due to diffraction. This enables scalable metasurface inverse design with near-linear complexity, facilitating the design of physically implementable and scalable, large-capacity DONNs.

Our main contributions can be summarized as follows:

- We analyze the core limitations of existing DONN training methods, which suffer from modeling inaccuracies and prohibitive simulation cost. We introduce SP^{2}RINT, a scalable PDE-constrained DONN optimization framework that ensures physical feasibility while drastically reducing simulation cost.
- **Spatially-Decoupled Metasurface Simulation**: SP^{2}RINT exploits the locality of meta-atom interactions to divide the metasurface into patches, enabling scalable, parallel simulation and reducing complexity from cubic to near-linear.
- **Progressive PDE-Constrained Learning**: SP^{2}RINT features a novel alternating scheme that interleaves relaxed DONN training with adjoint-based projection onto the Maxwell-constrained subspace, ensuring physical realizability throughout training while maintaining design flexibility and expressivity.
- On multiple DONN benchmarks, SP^{2}RINT achieves up to **63.88%** higher accuracy and delivers an **1825×** speed-up compared to state-of-the-art simulation-in-the-loop training methods because of the patched transfer matrices probing and much fewer iterations of inverse design, bridging the gap between analytical DONN training and physically realizable meta-optic hardware.

2 Background

This section introduces the physical principles of DONNs and existing DONN training methods.

2.1 Transfer Matrix Description of DONN Responses

Light propagation in DONNs involves two linear physical processes: free-space diffraction and light modulation, as illustrated in Fig. 1.

Free-Space Diffraction. Light diffraction in a homogeneous medium can be precisely modeled using a Green's function formulation, which provides an analytical near-to-far field transformation. This representation naturally involves Hankel functions to describe cylindrical wave propagation [1].

Metasurface-based Modulation. Metasurfaces modulate the phase and amplitude of light through subwavelength structures. The precise response of a metasurface is governed by Maxwell's equations.

In Fig. 1, a K-layer DONN, comprising alternating modulation and diffraction, is described by the Transfer Matrix Method (TMM), which naturally enables a layer-wise decomposition of wave propagation,

$$f(\mathbf{x}_{\text{in}};\boldsymbol{\epsilon}) = \left|\left(\prod_{i=1}^{K} U_i(z)\mathcal{T}_i(\epsilon_i)\right) U_0\mathbf{x}_{\text{in}}\right|^2 = |\mathbf{h}_{\text{out}}|^2. \tag{1}$$

where the inputs will be encoded to the intensity or phase of incident light $\mathbf{x}_{\text{in}}$, $U(z)$ is the light diffraction matrix over a distance z, and $\mathcal{T}(\epsilon)$ is the metasurface modulation matrix parameterized by its permittivity distribution ϵ. In prior work, based on LPA, a metasurface with n meta-atoms is modeled as an element-wise modulation plate with diagonal response $\mathcal{T} = \text{diag}([A_1 e^{j\phi_1}, \cdots, A_n e^{j\phi_n}])$. At the end, the light will be converted to electrical signals via photodetector arrays with a square function applied $|\cdot|^2$. The metasurface transfer matrix $\mathcal{T}(\epsilon)$ can be numerically obtained by stimulating the system with orthonormal basis vectors to capture its full impulse response.

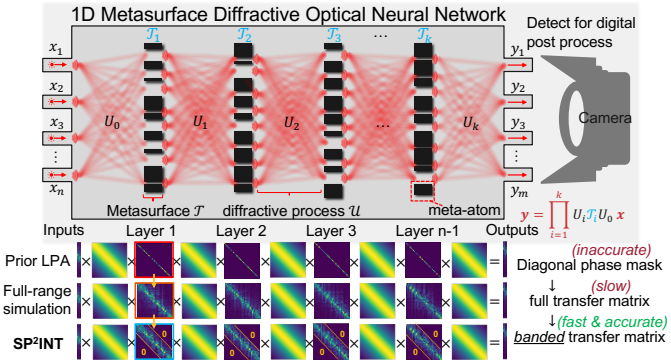


Figure 1: DONN can be modeled as cascaded transformation $\mathcal{T}$ and diffraction U. Metasurface transfer matrix modeling comparison among 3 methods. Our SP^{2}RINT uses banded transfer matrix probing for fast and accurate modeling.

Table 1: DONN training complexity. E denotes the number of training epochs, N is the size of the training dataset, B is the inverse design iteration budget for each training epoch ($B \ll N$), and n represents the spatial dimension of the metasurface.

DONN Training Methods		Algorithmic Complexity
Heuristics-based	LPA[20]	$O(EN+1)$
	Convolution[35]	$O(EN+1)$
Simulation-based	Simulation-in-the-loop[11]	$O(ENn^3)$
Simulation-based	SP^{2}RINT	$O(EN+EBn)$

2.2 Hybrid Diffractive Optical Neural Network

It is a promising trend to hybridize DONN with digital neural networks for adaptive AI inference [19, 34, 38]. Figure 1 shows a hybrid DONN with an optical feature extraction module and a lightweight digital neural head. As the light carrying the input length-n vector propagates through multiple metasurfaces, it forms an intensity distribution on the detector plane. This intensity pattern is coupled out as m-channel readout signals. This $m \times n$ linear unit can be used as a convolution with a single input channel, kernel size of n, and m output channels. The digital head can be trained for different downstream tasks.

2.3 Overview of DONN Training Methods

Heuristic DONN Training Methods. Heuristic approaches simplify the metasurface response, avoiding costly simulations during optimization, and implement the physical design post hoc [3–6, 8, 9, 12–16, 18, 20–22, 24–26, 30, 32, 33, 35, 36, 39, 41]. A widely used method models the metasurface $\mathcal{T}$ as a diagonal matrix with trainable phase shifts. However, this diagonal assumption neglects inter-element optical coupling, making the resulting $\mathcal{T}$ physically unrealizable.

Another method accounts for meta-atom interactions by modeling the metasurface as a learned convolution kernel [35], still relying on LPA. The reliance on LPA still limits accuracy, especially in regimes where strong inter-element interference or non-periodic layouts cause the approximation to break down.

Simulation-based Training Methods. To ensure physical fidelity, simulation-in-the-loop methods embed full-wave simulations directly into the DONN training process [11]. At each iteration, gradients are computed via the adjoint method [10]. While this approach provides accurate modeling and guarantees physical feasibility throughout training, its computational cost scales poorly with system size.

Table 1 compares the computational complexity of different DONN training paradigms in terms of training cost (related to E,N) and design cost (related to B,n). SP^{2}RINT offers a favorable trade-off by decoupling training from simulation. Its design cost scales linearly with metasurface size, thanks to the use of localized, patched transfer matrix probing. As a result, SP^{2}RINT retains physical fidelity while remaining computationally tractable.

2.4 Metasurface Design Methods

Various methods have been proposed for metasurface design. A widely adopted approach is the LUT-based method [2, 7], where phase elements are independently selected from a pre-simulated library/LUT and simply ensembled to approximate a target phase profile. Adjoint method-based inverse design [23, 28, 29, 40] directly optimizes the design variables in high-dimensional space to maximize a certain objective function using adjoint gradients. However, adjoint methods rely on computationally intensive full-wave optical simulations, which are prohibitively time-consuming. To clarify, DONN training is fundamentally more challenging than metasurface inverse design tasks. Inverse design solves a deterministic optimization problem for a specific figure-of-merit. In contrast, DONNs must learn metasurface designs on thousands of input-label pairs in the training dataset and generalize to the test set, making it a PDE-constrained learning task.

3 Proposed DONN Training Scheme SP^{2}RINT

We aim to train DONNs under physically realistic conditions that obey Maxwell's equations. We formulate DONN training as a PDE-constrained subspace optimization problem. To address both physical realizability and scalability challenges, we introduce SP^{2}RINT, a progressive training framework that alternates between relaxed DONN learning and adjoint-based inverse design. SP^{2}RINT integrates two key innovations: **progressive soft projection**, which gradually enforces physical constraints while enabling efficient design space exploration, and **patched transfer matrix probing**, which enables scalable modeling of large metasurfaces without full-system simulations.

3.1 Problem Formulation

Instead of being formulated as a deterministic optimization problem as in conventional inverse design, DONN training is formulated as a PDE-constrained learning problem:

$$\begin{aligned} \boldsymbol{\epsilon}^*, w^* &= \underset{\boldsymbol{\epsilon}, w}{\operatorname{argmin}}\, \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathcal{L}\left(g_w \circ f(\mathbf{x};\boldsymbol{\epsilon}), y\right)\right],\ f(\mathbf{x};\boldsymbol{\epsilon}) = |\mathbf{h}|^2, \\ &\text{s.t. } \left(\nabla \times (\boldsymbol{\epsilon}_{\text{tot}}^{-1}\nabla\times) - \omega^2\mu_0\epsilon_0\right)\mathbf{h} = \mathbf{b}(\mathbf{x}) \ \rightarrow\ A(\boldsymbol{\epsilon})\mathbf{h} = \mathbf{b}(\mathbf{x}). \end{aligned} \tag{2}$$

Here, $f(\mathbf{x};\boldsymbol{\epsilon})$ models the diffractive optical system, its output being the light intensity detected on photodetector arrays (for TM polarized light, this intensity is proportional to $|\mathbf{h}|^2$, where $\mathbf{h}$ is the magnetic field). The function $g_w(\cdot)$ represents the subsequent digital processing head, encompassing operations like normalization, biasing, and the final digital classification layers, with w as its learnable variables. The primary trainable parameters for the optical system are the k metasurface structures, characterized by their permittivity distributions $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_k)$. Input data and label pairs $(\mathbf{x}, y)$ are drawn from the training dataset $\mathcal{D}$, and $\mathcal{L}$ is the chosen loss function. The magnetic field $\mathbf{h}$ follows Maxwell's equation, which sets a challenging Maxwell PDE constraint to the optimization problem. $\boldsymbol{\epsilon}_{\text{tot}}$ describes the entire optical system, including multiple cascaded metasurfaces and their spacing cladding. For simplicity, we use a linear equation for the PDE constraint $A(\boldsymbol{\epsilon})\mathbf{h} = \mathbf{b}(\mathbf{x})$, where $\mathbf{h}$ is the vectorized magnetic field and $\mathbf{b}(\mathbf{x})$ is the vectorized input source related to $\mathbf{x}$.

This PDE-constrained formulation dictates that each iteration requires solving the forward Maxwell equation to obtain $\mathbf{h}$, followed by solving an adjoint Maxwell equation to compute the gradients $\frac{d\mathcal{L}}{d\boldsymbol{\epsilon}}$ with respect to the metasurface parameters:

$$A(\boldsymbol{\epsilon})^\top \mathbf{h}_{\text{adj}} = -\frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\mathbf{h}},\ \frac{\mathrm{d}\mathcal{L}}{\mathrm{d}\boldsymbol{\epsilon}} = -\Re(\mathbf{h}_{\text{adj}}^\top \mathbf{h}). \tag{3}$$

However, rigorous full-wave simulation of the entire optical system is prohibitively time-consuming, making it impractical to embed this simulation within the outer training loop.

3.2 Understanding Difficulty in Designing Implementable DONNs

3.2.1 The Essence of the PDE Constraint. As shown in Fig. 1 and Eq. (1), each metasurface with permittivity ϵ_i has a transfer matrix $\mathcal{T}_i(\epsilon_i)$ that maps the input $\mathbf{x}_i$ to the output field near the metasurface. However, not every arbitrary transfer matrix $\mathcal{T}$ can find its physical metasurface implementation. **Therefore, training a DONN that adheres to Maxwell's equations inherently becomes a constrained optimization problem, where the learned transfer matrices $\widehat{\mathcal{T}}_i(\epsilon_i)$ must reside in the subspace of physically implementable transformations**, which we refer to as the *implementable subspace*.

3.2.2 Training Difficulties. We identify **three key difficulties** that hinder the design of physically implementable and high-performance DONNs: ❶ **Characterizing the Valid Subspace:** The intricate optical physics make a precise analytical characterization of the implementable subspace unavailable, hence it is difficult to guide the optimization effectively. ❷ **Prohibitive Simulation Cost:** Directly enforcing physical realizability by repeatedly solving PDEs during training is computationally formidable. ❸ **Non-Convex Optimization Landscape:** The subspace of implementable transfer matrices $\mathcal{T}_i$ is highly non-convex, which significantly increases the risk of the training process prematurely converging to bad local optima. This, in turn, curtails effective exploration of the vast design space and impedes the discovery of high-expressivity metasurface designs.

To effectively address these challenges, several fundamental **questions** must first be answered:

Q1. How to effectively restrict transfer matrices within the implementable subspace? While *penalty methods* require unavailable analytical subspace characteristization and *reparametrization methods* require unaffordable PDE solving per iteration, we choose to use **progressive projection (i.e., periodically perform inverse design)** to balance efficiency and PDE constraint compliance.

Q2. How can we avoid per-iteration, unscalable PDE solving? While TMM allows us to decompose full-system simulation into layer-wise subproblems, each layer still requires expensive forward/adjoint simulations on the whole metasurface. To make training scalable, we must **decouple field evaluation $\mathbf{h}_i$ from design updates on ϵ_i and solve PDEs only when necessary.**

Q3. How to perform projection to balance exploration and efficiency? Since metasurface inverse design is inherently a slow and difficult binary optimization problem, effective training requires a carefully scheduled projection strategy. *Hard projection* strictly enforces feasibility but limits exploration and increases simulation cost, while *soft projection* improves efficiency and optimization flexibility but risks constraint violation. **To ensure fast convergence and maintain physical realizability, we need to balance these trade-offs through progressive projection that gradually tightens constraints during training.**

3.3 Overview of Proposed DONN Training Flow

To efficiently solve Eq. (2), we propose a *variable separation* method to decouple the transfer matrix $\widehat{\mathcal{T}}$ used in DONN training from the

Algorithm 1: Progressive Projected Training for DONN

Input: DONN model $\mathcal{M}_{\widehat{\mathcal{T}},w}$, use $\mathcal{M}$ for short; Training set $\mathcal{D}(x,y)$; Uniform metasurface init ϵ^0; Initial and final binarization sharpness: s_0, s_T; Inverse design iteration budget per epoch B, number of inverse design iteration per projection I; Number of metasurface layers K; patch size P

Output: Trained model with implementable metasurfaces

1 $\boldsymbol{\epsilon}^0 \leftarrow [\epsilon^0] \times K$; $n_p \leftarrow B/I$; $\mathcal{S} \leftarrow \texttt{Sched}(s_0, s_T, B)$
2 **for** *epoch* $t \leftarrow 1$ **to** n_{epochs} **do** // Sec. 3.3
3 $\boldsymbol{\epsilon}^t \leftarrow \boldsymbol{\epsilon}^{t-1}$;
4 $\widehat{T}^t \leftarrow \texttt{ProbeTM}(\epsilon^t, s_T, P)$; // Sec. 3.4
5 $\mathcal{M}.\texttt{SetTM}(\widehat{T}^t, k), \forall k \in [K]$; // Load $\mathcal{T}$ for k^{th} layer
6 **foreach** *batch index* j, (x,y) *in* $\mathcal{D}$ **do**
7 $\texttt{TrainStep}(\mathcal{M}, \text{x}, \text{y})$; // Train $\mathcal{M}$
8 **if** $j \bmod b_p = 0$ ***and*** $j \neq 0$ **then**
9 $\epsilon_i^t \leftarrow \texttt{Proj}(\widehat{\mathcal{T}}_i), \forall i \in [K]$; // Solve Eq.(6)
10 $\boldsymbol{\epsilon}^t \leftarrow \texttt{Proj}(\widehat{\mathcal{T}}_{\text{tot}}$ of $\mathcal{M})$; // Solve Eq. (6) and (7)
11 $\mathcal{S}.\texttt{reset}()$; // Sec. 3.5

metasurface response $\mathcal{T}$ that requires n *full-metasurface* simulations to extract. The formulation is rewritten as

$$\widehat{\mathcal{T}}^*, w^* = \underset{\widehat{\mathcal{T}},w}{\operatorname{argmin}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathcal{L}\left(g_w \circ f(\mathbf{x}; \widehat{\mathcal{T}}), y\right)\right], \quad \text{s.t. } A(\epsilon_i)\mathcal{T}_i = B,\ \forall i \in [K];\ \widehat{\mathcal{T}}_i = \mathcal{T}_i,\ \forall i \in [K]. \tag{4}$$

To solve this constrained optimization, the projected gradient descent can be adopted, as shown in Alg. 1, where we alternatively solve two subproblems. The first subproblem is ❶ **Relaxed DONN Training**:

$$\widehat{\mathcal{T}}^{t+1}, w^{t+1} = \underset{\widehat{\mathcal{T}},w}{\operatorname{argmin}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \left[\mathcal{L}\left(g_{w^t} \circ f(\mathbf{x}; \widehat{\mathcal{T}}), y\right)\right]. \tag{5}$$

The second subproblem is ❷ **Metasurface Inverse Design**: optimizing the metasurface permittivity $\boldsymbol{\epsilon}$ to match the target matrix $\widehat{\mathcal{T}}$:

$$\boldsymbol{\epsilon}^{t+1} = \underset{\boldsymbol{\epsilon}}{\operatorname{argmin}} \sum_{i=1}^{K} \|\mathcal{T}_i - \widehat{\mathcal{T}}_i^{t+1}\|_F^2, \quad \text{s.t. } A(\epsilon_i)\mathcal{T}_i = B,\ \forall i \in [K]. \tag{6}$$

This subproblem is very time-consuming. The following section discusses how we efficiently solve ❷.

3.4 Spatially-Decoupled Metasurface Simulation for Scalable Transfer Matrix Probing

In subproblem ❷, we need to extract $\mathcal{T}_i$ for all k metasurfaces by solving nk full-metasurface PDEs in total. Guided by the physics prior of short-range meta-atom interactions, we observe a **banded diagonal** structure in $\mathcal{T}$. The light shone on one meta-atom will not scatter to far-away near-field locations, allowing us to simulate only a small local region and ignore the exterior with acceptable error in Fig. 2. Hence, we propose a spatially-decoupled approach by cutting a large metasurface into overlapping patches of P meta-atoms with stride 1. For a metasurface with n meta-atoms, we only need to simulate n small size-P patches, greatly reducing transfer matrix extraction cost. The cost **scales linearly with metasurface size** under sequential simulation, and can approach constant runtime under parallel simulation because the patches are fully decoupled, replacing cubic complexity with a far more scalable approach.

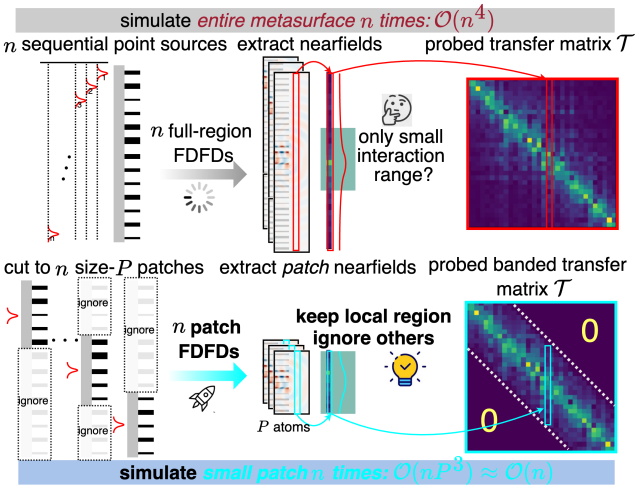


Figure 2: Proposed spatially-decoupled transfer matrix probing method cuts the metasurface into small patches for patch simulation that reduces complexity from cubic to linear.

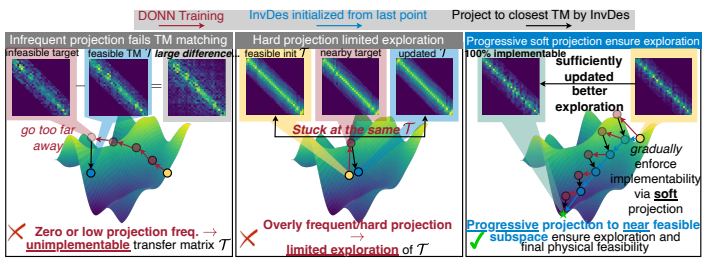


Figure 3: Our proposed SP²RINT framework enables both exploitation and physical feasibility.

3.5 Progressive PDE-Constrained Training

It is critical to carefully schedule the alternating frequency between ❶ and ❷ as well as the binarization sharpness in ❷. As shown in Fig. 3, a full epoch of unconstrained training often causes $\widehat{\mathcal{T}}_i$ to drift too far from the subspace, such that the subsequent projection step can no longer find a physically realizable metasurface that accurately recovers the target response. This results in a severe performance drop and even divergence. Increasing the projection frequency helps prevent $\widehat{\mathcal{T}}_i$ from drifting too far from the subspace. However, due to the highly non-convex nature of the subspace, as discussed in Sec. 3.2.2, overly frequent hard projection often causes the optimization to get stuck near the initial point without meaningful progress, shown in Fig. 3. To resolve this, we introduce a *progressive PDE-constrained training* strategy. Rather than enforcing hard binarization from the start, we allow the inverse design module to initially project to relaxed, continuous-valued patterns that lie near the subspace. As training proceeds, we gradually tighten the binary constraint, ensuring the final metasurfaces are both expressive and physically realizable.

3.6 System Fine-Tuning for Enhanced Optimality

As shown in Alg. 1 Line 10, besides matching $\mathcal{T}_i$ of each individual metasurface (Eq. (6)), we introduce an extra fine-tuning stage to further calibrate the response of the entire optical system. This is achieved by optimizing the metasurface ϵ to match the end-to-end system-level transfer matrix,

$$\epsilon^{t+1} = \arg\min_{\epsilon} \left\| \mathcal{T}_{\text{tot}} S - \widehat{\mathcal{T}}_{\text{tot}}^{t+1} S \right\|_F^2, \quad \mathcal{T}_{\text{tot}} = \prod_{i=1}^{K} U_i \mathcal{T}_i(\epsilon_i), \quad \text{s.t.} \quad A(\epsilon_i)\mathcal{T}_i = B, \quad \forall i \in [K]. \tag{7}$$

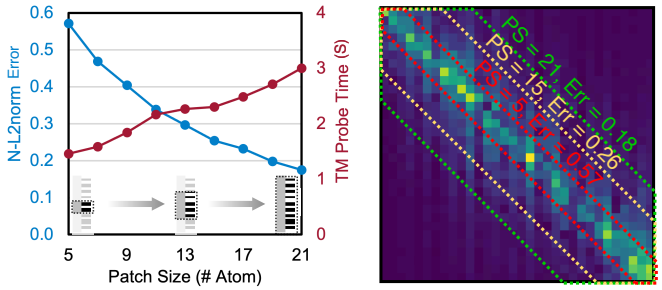


Figure 4: Different patch size trades off transfer matrix probing error and runtime.

$\mathcal{T}_{\text{tot}}$ is the total transfer matrix of the multi-layer DONN. The target total transfer matrix, $\widehat{\mathcal{T}}_{\text{tot}}^{t+1}$, is constructed from the layer-wise target modulation matrices $\widehat{\mathcal{T}}_i^{t+1}$ obtained from the DONN training stage shown in Eq. (5), as $\widehat{\mathcal{T}}_{\text{tot}}^{t+1} = \prod_{i=1}^{K} U_i \widehat{\mathcal{T}}_i^{t+1}$. The matrix S consists of a set of probing bases. This system fine-tuning enables improved alignment with the desired response and enhances DONN performance.

4 Evaluation

4.1 Evaluation Settings

The hybrid DONN for classification comprises an optical feature extractor and a digital head. The optical extractor has 2 cascaded metasurfaces and operates as a 3×3 convolution (9 in-ports) with square nonlinearity, a single input channel, and 4 output channels. RGB images are converted to grayscale, and pixels are phase encoded from [0-1] to $[0,\pi]$. For simplicity, we adopt a 1D metasurface design, where each metasurface consists of 32 Si meta-atoms with air cladding, a period of 300 nm and a pillar height of 750 nm. The pillar widths are trainable. The diffraction distances z between layers are set to 4 μm. The wavelength is 850 nm. The digital head is a 4-channel 1×1 convolution followed by a pooling layer and 2 linear layers.

4.2 Patch Size Selection: Efficiency vs. Accuracy

We first determine a critical hyperparameter: the patch size in patched TM probing. In Fig. 4, larger patch sizes improve the fidelity of transfer matrix probing by capturing more near-field interactions, but this comes at the cost of increased simulation time. To balance probing accuracy and efficiency, we set the patch size to 17 meta-atoms in all experiments. Note that the optimal patch size depends on factors such as the wavelength, the meta-atom material, and the diffraction distance between the metasurface and the observation plane.

4.3 Main Results

Table 2 compares different DONN training methods. Fashion-MNIST [37] and SVHN [27] represent image recognition (metric is *accuracy and cross-entropy (CE) loss*), while Darcy Flow [17] shows PDE solving tasks (metric is *normalized L2-Norm*). We evaluate the test set performance using the final trained DONNs with real simulated responses of implemented metasurfaces. We compare SP^{2}RINT with three heuristic methods (LPA, conv LPA, and smoothed) and simulation-in-the-loop training. The *Smoothed Metasurface* method uses a regularization term to enforce similar meta-atom sizes across neighbors [41].

To clarify, train-test performance gap is mainly due to *metasurface modeling error* and *NN generalization error*. Since all heuristic methods oversimplify the complex metasurface responses, the optimized transfer matrices are *not physically realizable*, hence they exhibit significant performance degradation when mapped to real metasurfaces during inference. The huge drop is mostly due to inaccurate metasurface modeling. In contrast, the optimized transfer matrices produced by

Table 2: Comparison across different DONN training methods on different benchmarks, our proposed SP^{2}RINT consistently achieves best performance. *Sim-in-the-loop* method has low accuracy after epoch 1 and takes 2 years to finish 100 epochs.

Benchmark	Baselines	Train CE	Train Acc	Test CE	Test Acc
Fashion-MNIST[37]	SP^{2}RINT	9.00E-02	96.85%	0.45	88.44%
	LPA[20]	1.52E-01	94.77%	2.25	58.79%
	conv LPA [35]	1.69E-01	94.11%	6.40	13.45%
	smoothed metasurface [14]	2.39E-02	99.45%	22.61	16.21%
	sim-in-the-loop [11]	Time out	Time out	Time out	Time out
avg. improv.		+66.67%			
SVHN[27]	SP^{2}RINT	3.16E-01	91.03%	0.76	81.61%
	LPA[20]	2.99E-01	91.70%	6.71	23.92%
	conv LPA [35]	2.21E-01	93.67%	2.51	11.51%
	smoothed metasurface [14]	5.98E-02	98.55%	5.04	7.23%
	sim-in-the-loop [11]	Time out	Time out	Time out	Time out
avg. improv.		+82.58%			
		Train N-L2norm		Test N-L2norm	
Darcy Flow[17]	SP^{2}RINT	0.37		0.35	
	LPA[20]	0.32		0.60	
	conv LPA [35]	0.33		0.87	
	smoothed metasurface [14]	0.33		0.48	
	sim-in-the-loop [11]	Time out		Time out	
avg. improv.		+42.40%			
total avg. improv.		+63.88%			

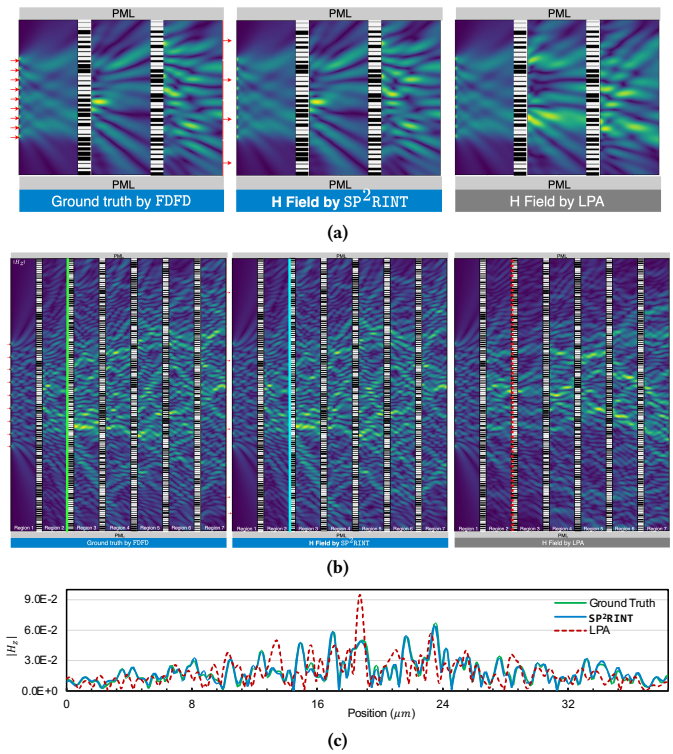


Figure 5: $|H_z|$ fields comparison on the 9-in-4-out metasurface system. (a) 2-layer 32-metaatom, (b) 6-layer 128-metaatom, (c) Magnetic field H_z slice comparison. Our SP^{2}RINT can accurately capture the metasurface transfer matrix, much more accurate than simple phase mask modeling using LPA, yielding almost the same H field amplitude calculated by FDFD.

SP^{2}RINT are guaranteed to lie within the implementable subspace. Hence, **the training accuracy evaluated using $\widehat{\mathcal{T}}$ remains almost the same when mapped to real metasurface responses $\mathcal{T}(\epsilon)$.** Our method eliminates the metasurface modeling error (our main goal) and achieves the best test accuracy up to the inevitable train-test generalization gap (not the issue we aim to solve in this work). Figure 5 visually compares the estimated and simulated optical fields within the diffractive feature extractor during inference. SP^{2}RINT accurately reproduces the field calculated by FDFD simulation, whereas LPA shows a significant field approximation error.

Table 3: Compare different projection sharpness schedules. Progressively tightening the projection from a soft to a hard-binarization device ensures both exploration and physical feasibility. *CE* is Cross-Entropy.

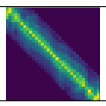
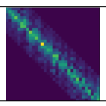
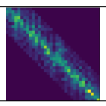
Sharpness Schedule	Per Projection	**Per Epoch**	Per Train
CE Loss ↓	0.579	**0.428**	0.954
Accuracy ↑	86.19%	**88.32%**	80.82%
Transfer Matrix			

Table 4: Comparison between the projection schedule in one training epoch. By default, we define 1 unit budget as 20 iterations. Each iteration includes 2 times of forward transfer matrices probing and corresponding adjoint simulation.

Cost	Proj. frequency	Proj. iters	CE ↓	Accuracy ↑
budget ×1	1 / epoch	20	1.190	56.04%
	2 / epoch	10	0.939	71.68%
	4 / epoch	5	0.360	87.59%
	5 / epoch	4	0.444	84.99%
	10 / epoch	**2**	0.352	89.54%
	20 / epoch	1	0.354	88.64%
budget ×1/2	5 / epoch	2	0.514	87.64%
budget ×1/2	10 / epoch	1	0.406	88.76%
budget ×2	10 / epoch	4	0.452	83.91%

For the simulation-in-the-loop method [11], we estimate the runtime for a single epoch to be 50k/32 × (2 metasurfaces) × (32 sims/TMProb) × 6.4s/sim = ∼ 178 hours for Fashion-MNIST with batch size 32, due to the large number of simulations involved, making it prohibitively time-consuming. It needs 2 years to finish 100 epochs. We consider this intractable and do not report results for this baseline. SP^{2}RINT only takes 5.85 min per epoch, which directly shows 1825× speedup.

4.4 Discussion on Projection Configurations

4.4.1 Progressive Soft Projection Schedule. As shown in Table 3, when strict binarization is enforced at every projection step, the transfer matrix of metasurfaces stagnates with minimal updates compared to initialization. If the binarization is only enforced at the end of the training process, it over-relaxes the inverse design process (i.e., projection) and fails to adjust model weights to tolerate the permittivity binarization effects, thus leading to low accuracy after convergence. Our proposed progressive soft projection gradually increases the binarization sharpness over time, allowing the metasurface to evolve meaningfully and contribute to higher classification accuracy, while still ensuring implementability through the final binarization.

4.4.2 Projection Frequency. Table 4 compares different projection frequencies under different runtime budgets. Increasing the frequency while proportionally reducing the number of iterations per projection maintains the overall runtime but significantly improves test accuracy. In our setting, we use 20 inverse design iterations per epoch with 10 projections per epoch, i.e., 2 adjoint gradient updates per projection.

4.4.3 Inverse Design Projection Objective. Figure 6 shows how different projection objectives impact the performance drift from unconstrained target performance to the feasible performance. Layer-wise projection with our extra system-level fine-tuning effectively mitigates the performance drift caused by the inverse design imperfection.

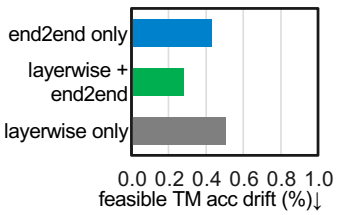


Figure 6: Layerwise+end-to-end inverse design gives the best projection accuracy.

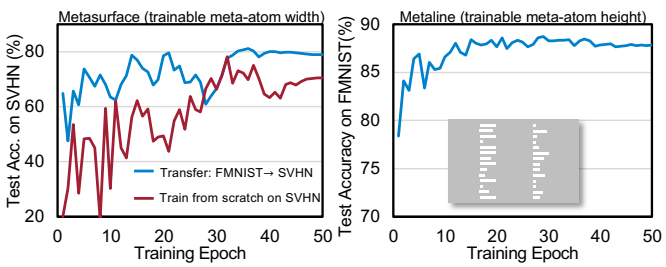


Figure 7: (a) Transfer learning of DONNs from Fashion-MNIST to SVHN shows faster convergence and higher performance than training on SVHN from scratch. (b) SP^{2}RINT can be generally applied to other devices, such as metalines with trainable meta-atom height.

4.4.4 Transfer Learning from Fashion-MNIST to SVHN. To demonstrate the effectiveness of the learned optical feature extractor, we performed transfer learning from Fashion-MNIST to SVHN. Despite the visual differences between the datasets, one consisting of grayscale images of clothing and the other of colorful house numbers, the transferred model shows faster convergence and improved performance compared to training from scratch, as shown in Fig. 7(a).

4.4.5 Generalization to Other Types of Meta-Optic Structures. Beyond metasurfaces, where light modulation is realized by varying the width of meta-atoms with fixed height, alternative approaches exist to control light behavior. For example, metalines achieve light modulation by fixing the width of meta-atoms and varying their height. Our proposed SP^{2}RINT framework is compatible with such architectures and can be used to train DONNs implemented with metalines. This is demonstrated in Fig. 7(b), which shows the training dynamics of a metaline-based DONN.

5 Conclusion

We propose SP^{2}RINT, a scalable and physically grounded framework for training DONN. SP^{2}RINT combines PDE-constrained optimization with spatially decoupled transfer matrix probing and progressive projection to ensure physical realizability without requiring per-iteration full-wave simulations. Evaluation shows that SP^{2}RINT achieves average test accuracy improvements of 63.88% over widely used heuristic training methods and delivers 1825× speed-up compared to the simulation-in-the-loop training. These results highlight SP^{2}RINT's potential to bridge the gap between high-performance deep learning and physically implementable nanophotonic hardware, enabling scalable, generalizable, and deployable optical neural systems.

6 Acknowledgment

The authors would like to acknowledge the NVIDIA Academic Grant Program Award.

References

[1] 2025. Flexcompute Tidy3D, https://github.com/flexcompute/tidy3d. (2025).

[2] Francesco Aieta, Mikhail A Kats, Patrice Genevet, and Federico Capasso. 2015. Multiwavelength achromatic metasurfaces by dispersive phase compensation. *Science* 347, 6228 (2015), 1342–1345.

[3] Bijie Bai, Xilin Yang, Tianyi Gan, Jingxi Li, Deniz Mengu, Mona Jarrahi, and Aydogan Ozcan. 2024. Pyramid diffractive optical networks for unidirectional image magnification and demagnification. *Light: Science & Applications* 13, 1 (2024), 178.

[4] Hang Chen, Jianan Feng, Minwei Jiang, Yiqun Wang, Jie Lin, Jiubin Tan, and Peng Jin. 2021. Diffractive deep neural networks at visible wavelengths. *Engineering* 7, 10 (2021), 1483–1491.

[5] Ruiyang Chen, Yingjie Li, Minhan Lou, Jichao Fan, Yingheng Tang, Berardi Sensale-Rodriguez, Cunxi Yu, and Weilu Gao. 2022. Physics-Aware Machine Learning and Adversarial Attack in Complex-Valued Reconfigurable Diffractive All-Optical Neural Network. *Laser Photonics Rev.* (2022).

[6] Ruiyang Chen, Yingjie Li, Minhan Lou, Cunxi Yu, and Weilu Gao. 2022. Complex-valued reconfigurable diffractive optical neural networks using cost-effective spatial light modulators. In *CLEO: Applications and Technology*. Optica Publishing Group, JTh3B–56.

[7] Wei Ting Chen, Alexander Y Zhu, Vyshakh Sanjeev, Mohammadreza Khorasaninejad, Zhujun Shi, Eric Lee, and Federico Capasso. 2018. A broadband achromatic metalens for focusing and imaging in the visible. *Nature nanotechnology* 13, 3 (2018), 220–226.

[8] Shane Colburn, Yi Chu, Eli Shilzerman, and Arka Majumdar. 2019. Optical frontend for a convolutional neural network. *Applied optics* 58, 12 (2019), 3179–3186.

[9] Ze Gu, Qian Ma, Xinxin Gao, Jian Wei You, and Tie Jun Cui. 2024. Direct electromagnetic information processing with planar diffractive neural network. *Science Advances* 10, 29 (2024), eado3937.

[10] Tyler W. Hughes, Momchil Minkov, Ian A. D. Williamson, and Shanhui Fan. 2018. Adjoint Method and Inverse Design for Nonlinear Nanophotonic Devices. *ACS Photonics* 5, 12 (Dec. 2018), 4781–4787. doi:10.1021/acsphotonics.8b01523 Addition/Correction published December 3, 2018.

[11] Erfan Khoram, Ang Chen, Dianjing Liu, Lei Ying, Qiqi Wang, Ming Yuan, and Zongfu Yu. 2019. Nanophotonic media for artificial neural inference. *Photonics Research* 7, 8 (2019), 823–827.

[12] Jingxi Li, Yi-Chun Hung, Onur Kulce, Deniz Mengu, and Aydogan Ozcan. 2022. Polarization multiplexed diffractive computing: all-optical implementation of a group of linear transformations through a polarization-encoded diffractive network. *Light: Science & Applications* 11, 1 (2022), 153.

[13] Yingjie Li, Ruiyang Chen, Weilu Gao, and Cunxi Yu. 2022. Physics-aware differentiable discrete codesign for diffractive optical neural networks. In *Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design*. 1–9.

[14] Yingjie Li, Ruiyang Chen, Minhan Lou, Berardi Sensale-Rodriguez, Weilu Gao, and Cunxi Yu. 2024. LightRidge: An End-to-end Agile Design Framework for Diffractive Optical Neural Networks. In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 4* (Vancouver, BC, Canada) *(ASPLOS '23)*. Association for Computing Machinery, New York, NY, USA, 202–218. doi:10.1145/3623278.3624757

[15] Yingjie Li, Weilu Gao, and Cunxi Yu. 2023. Rubik's optical neural networks: multi-task learning with physics-aware rotation architecture. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (Macao, P.R.China) *(IJCAI '23)*. Article 847, 10 pages. doi:10.24963/ijcai.2023/847

[16] Yuhang Li, Yi Luo, Bijie Bai, and Aydogan Ozcan. 2023. Analysis of Diffractive Neural Networks for Seeing Through Random Diffusers. *IEEE Journal of Selected Topics in Quantum Electronics* 29, 2: Optical Computing (2023), 1–17. doi:10.1109/JSTQE.2022.3194574

[17] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. 2021. Fourier Neural Operator for Parametric Partial Differential Equations. In *Proc. ICLR*.

[18] Xing Lin, Yair Rivenson, Nezih T Yardimci, Muhammed Veli, Yi Luo, Mona Jarrahi, and Aydogan Ozcan. 2018. All-optical machine learning using diffractive deep neural networks. *Science* 361, 6406 (2018), 1004–1008.

[19] Wencan Liu, Yuyao Huang, Run Sun, Tingzhao Fu, Sigang Yang, and Hongwei Chen. 2025. Ultra-compact multi-task processor based on in-memory optical computing. *Light: Science & Applications* 14, 1 (2025), 134.

[20] Xuhao Luo, Yueqiang Hu, Xiangnian Ou, Xin Li, Jiajie Lai, Na Liu, Xinbin Cheng, Anlian Pan, and Huigao Duan. 2022. Metasurface-enabled on-chip multiplexed diffractive neural networks in the visible. *Light: Science & Applications* 11, 1 (2022), 158.

[21] Yi Luo, Deniz Mengu, Nezih T Yardimci, Yair Rivenson, Muhammed Veli, Mona Jarrahi, and Aydogan Ozcan. 2019. Design of task-specific optical systems using broadband diffractive neural networks. *Light: Science & Applications* 8, 1 (2019), 112.

[22] Yi Luo, Yifan Zhao, Jingxi Li, Ege Çetintaş, Yair Rivenson, Mona Jarrahi, and Aydogan Ozcan. 2022. Computational imaging without a computer: seeing through random diffusers at the speed of light. *eLight* 2, 1 (2022), 4.

[23] Mahdad Mansouree, Andrew McClung, Sarath Samudrala, and Amir Arbabi. 2021. Large-scale parametrized metasurface design using adjoint optimization. *Acs Photonics* 8, 2 (2021), 455–463.

[24] Deniz Mengu and Aydogan Ozcan. 2022. All-optical phase recovery: diffractive computing for quantitative phase imaging. *Advanced Optical Materials* 10, 15 (2022), 2200281.

[25] Deniz Mengu, Anika Tabassum, Mona Jarrahi, and Aydogan Ozcan. 2023. Snapshot multispectral imaging using a diffractive optical network. *Light: Science & Applications* 12, 1 (2023), 86.

[26] Deniz Mengu, Muhammed Veli, Yair Rivenson, and Aydogan Ozcan. 2022. Classification and reconstruction of spatially overlapping phase images using diffractive optical networks. *Scientific Reports* 12, 1 (2022), 8446.

[27] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, et al. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *Proc. NIPS*.

[28] Jaewon Oh, Kangmei Li, Jun Yang, Wei Ting Chen, Ming-Jun Li, Paulo Dainese, and Federico Capasso. 2022. Adjoint-optimized metasurfaces for compact mode-division multiplexing. *ACS Photonics* 9, 3 (2022), 929–937.

[29] Thaibao Phan, David Sell, Evan W Wang, Sage Doshay, Kofi Edee, Jianji Yang, and Jonathan A Fan. 2019. High-efficiency, large-area, topology-optimized metasurfaces. *Light: Science & Applications* 8, 1 (2019), 48.

[30] Chao Qian, Xiao Lin, Xiaobin Lin, Jian Xu, Yang Sun, Erping Li, Baile Zhang, and Hongsheng Chen. 2020. Performing optical logic operations by a diffractive neural network. *Light: Science & Applications* 9, 1 (2020), 59.

[31] Damián Rodríguez-Trujillo, Alicia E. Torres-García, Mikel Aldea, Jorge Teniente, Asier Marzo-Pérez, and Miguel Beruete. 2025. Phase Smoothing for Mutual Coupling Mitigation in Multifunctional Metasurfaces Designed with Diffractive Neural Networks. *Advanced Optical Materials* (2025), e02746.

[32] Yingheng Tang, Ruiyang Chen, Minhan Lou, Jichao Fan, Cunxi Yu, Andy Nonaka, Zhi Yao, and Weilu Gao. 2025. Optical Neural Engine for Solving Scientific Partial Differential Equations. *Nature Communications*.

[33] Ethan Tseng, Shane Colburn, James Whitehead, Luocheng Huang, Seung-Hwan Baek, Arka Majumdar, and Felix Heide. 2021. Neural nano-optics for high-quality thin lens imaging. *Nature communications* 12, 1 (2021), 6493.

[34] Kaixuan Wei, Xiao Li, Johannes Froech, Praneeth Chakravarthula, James Whitehead, Ethan Tseng, Arka Majumdar, and Felix Heide. 2024. Spatially varying nanophotonic neural networks. *Science Advances* 10, 45 (2024), eadp0391.

[35] Zhicheng Wu, Ming Zhou, Erfan Khoram, Boyuan Liu, and Zongfu Yu. 2019. Neuromorphic metasurface. *Photonics Research* 8, 1 (2019), 46–50.

[36] Jinlin Xiang, Shane Colburn, Arka Majumdar, and Eli Shlizerman. 2022. Knowledge distillation circumvents nonlinearity for optical convolutional neural networks. *Applied Optics* 61, 9 (2022), 2173–2183.

[37] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *Arxiv* (2017).

[38] Ziang Yin, Yu Yao, Jeff Zhang, and Jiaqi Gu. 2025. CHORD: Composable Hybrid Optical Reconfigurable Diffractive Framework for Optical Neural Network. In *Proc. DAC*.

[39] Haoyi Yu, Zihao Huang, Simone Lamon, Baokai Wang, Haibo Ding, Jian Lin, Qi Wang, Haitao Luan, Min Gu, and Qiming Zhang. 2025. All-optical image transportation through a multimode fibre using a miniaturized diffractive neural network on the distal facet. *Nature Photonics* (2025), 1–8.

[40] Dasen Zhang, Zhenzhen Liu, Xiaotong Yang, and Jun Jun Xiao. 2022. Inverse design of multifunctional metasurface based on multipole decomposition and the adjoint method. *ACS Photonics* 9, 12 (2022), 3899–3905.

[41] Shanglin Zhou, Yingjie Li, Minhan Lou, Weilu Gao, Zhijie Shi, Cunxi Yu, and Caiwen Ding. 2023. Physics-aware roughness optimization for diffractive optical neural networks. In *Proc. DAC*. 1–6.